Juwon (Lucia) Park 1008754224
Cindy Yu 1009099431
Yuehan Li 1005778847

Host Order Effect on Virus Segmentation

Abstract

This study investigates whether mammalian host order significantly influences susceptibility to segmented viruses. Using generalized linear models and data from the Clover database, we compared models with and without host order as a predictor. Contrary to our prediction, host order did not emerge as a consistent determinant of segmentation susceptibility. These results challenge assumptions about the relationship between host taxonomy and segmented virus evolution, suggesting that other viral traits may play a more crucial role in susceptibility patterns.

Introduction

Viruses display an extraordinary range of characteristics that influence their ability to infect hosts across species. Among these, segmented viruses stand out due to their genomes being divided into multiple parts, which confer evolutionary advantages such as reassortment and increased control of gene expression. Reassortment, in particular, allows for viral progeny with segments from multiple parents (McDonald et al., 2016). This unique genomic architecture allows segmented viruses to persist in and transition between different ecological niches, posing significant challenges for predicting and managing emerging infectious diseases.

Host order has been proposed as a potential factor influencing the susceptibility of species to segmented viruses. For example, shared traits among host orders, such as receptor availability, immune response, or life history traits, may drive the evolution of segmented viral genomes. Based on these considerations, we hypothesize that host order plays a critical role in shaping the likelihood of infection by segmented viruses. This study aims to test this hypothesis using comprehensive host-virus interaction data to compare two logistic regression models, one with host order and one without.

Method

Data Description

We retrieved data from the Clover database, a resource on host-virus associations developed from four major datasets (HP3, GMPD2, EID2 and Shaw et al's) by Rory Gibb and collaborators, all edited to have taxonomy consistent with NCBI. This study focuses on segmented viruses, using two datasets: Mammal_viruses_Associations (mammalian host orders and their associated viruses) and HP3_virus (segmented genome, virus family, and average genome size).

For data preparation, virus names were standardized: replacing spaces with underscores, converting text to lowercase, and changing binary data to 0s and 1s. Columns were subsetted to include predictors hypothesized to influence segmentation: genome length, cytoplasmic replication, enveloped status, and single- or double-stranded. Overall, predictors were chosen based on their ability to change virus transmission and virulence. Single-stranded viruses generally mutate faster, which could result in a better ability to infect novel species (Sanjuán & Domingo-Calap, 2016). Cytoplasm-replicating virions organize their genomes and invade host defences using organelle-like structures, which is very different from replication in the nucleus and must be considered (den Boon et al., 2010). Longer genomes may benefit from segmentation to better regulate gene expression, but may also suffer from a lower mutation rate (Sanjuán & Domingo-Calap, 2016). Enveloped viruses have been found to more frequently cross species barriers and cause zoonotic infections (Valero-Rello & Sanjuán, 2022). Datasets will then be merged based on virus names, with rows containing missing values, and host orders with less than 40 data points removed.

Data analysis

We predict that host order affects whether a mammal species has a greater proportion of infection by segmented viruses relative to non-segmented viruses. Two generalized linear models (GLMs) following a binomial distribution will be tested against the data. Both models will regress the proportion of viruses that are segmented against multiple predictor variables. Only one of the two models will have host order as a predictor variable. Then, both models will be evaluated on their compatibility with the data. For our prediction to be correct, the GLM including the host order predictor will fit the data better and have lower AIC.

Our data follows a binomial distribution because we can treat segmentation as 'success'—i.e. a segmented virus is successful and equals 1, while a non-segmented virus is unsuccessful and equals 0. Since we are interested in host order, rather than host species, we can use each mammal species-virus association as a trial.

Results

Using our described method, we analyzed the relationship between host order and susceptibility to segmented viruses. Two logistic regression models (GLMs) were created to assess the predictors of virus segmentation. In terms of AIC, the model incorporating host order did outperform the model excluding it, indicating that host order is a driver of segmentation patterns. However, in examining the p-values, it seemed that host order was only sometimes significant. For instance, certain host orders (Chiroptera, Carnivora, Primates, Rodents, and Perissodactyla) showed a high degree of association (p-values $< 1*10^{-10}$) with the presence of segmented viruses. Other host orders were not significant on the $\alpha = 0.05$ level.

Discussion

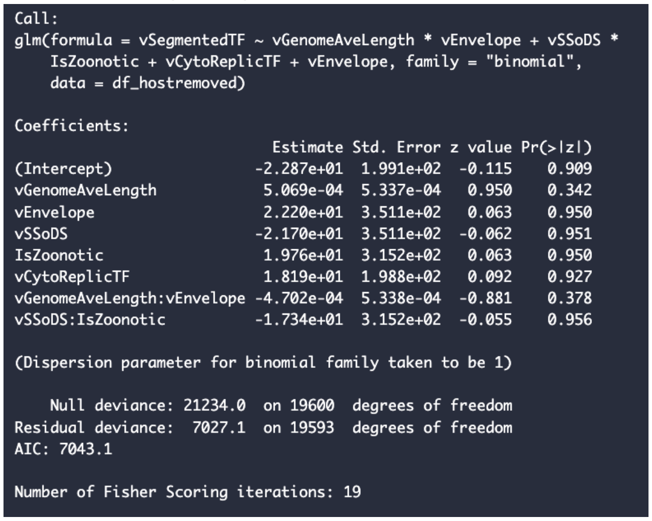
In conclusion, although the host order-inclusive AIC outperformed the null model, host order is, at best, sometimes a driver of infection by segmented viruses. Interestingly, out of significant host orders, all but Carnivora have negative estimates, implying that belonging to one of these host orders decreases the log-odds of an associated virus being segmented. This may be surprising, as rodents, primates, and bats are known to be common zoonotic reservoirs, and would be expected to be more affected by segmented viruses. Investigating the relationship between these zoonotic reservoirs, segmented viruses, and zoonotic viruses in general could elucidate the reasons for this surprising result.

The reasons behind specific orders having significant results are outside the scope of this analysis. Our model is limited by the chosen predictors; a more comprehensive model would benefit from determining collinear predictors and traits most affecting virus segmentation. In the future, the data exploration could include a Principal Components Analysis to gauge which traits explain the most variance in the data. This analysis was also limited by the dataset and sampling bias. There is likely more information on zoonotic viruses. Furthermore, after data wrangling, the number of rows in our dataset was reduced from over 50 thousand to approximately 20 thousand, implying that a large amount of data is missing.

Supplemental

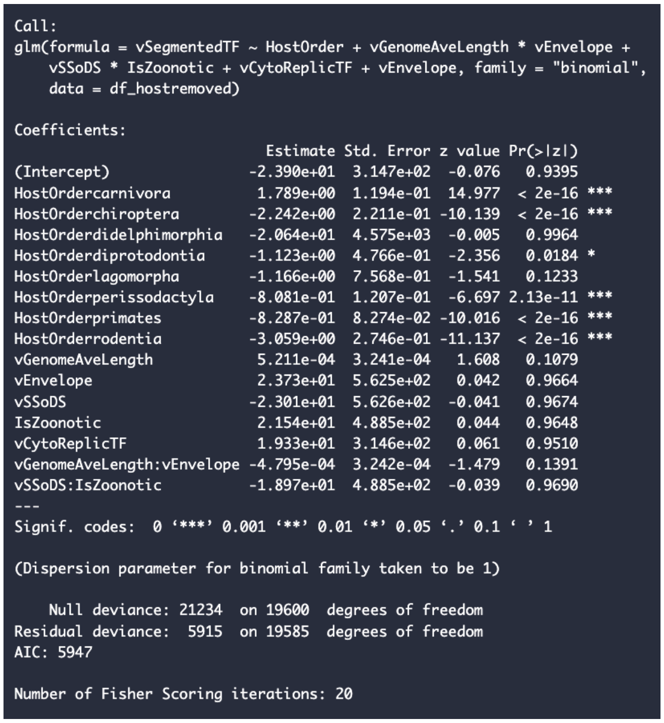
Column information is in the README.txt.
Final dataset is called FINALDATASET.csv.
Code is in GroupProject_Segmented.rmd

The following images are the summaries of the created GLMs.



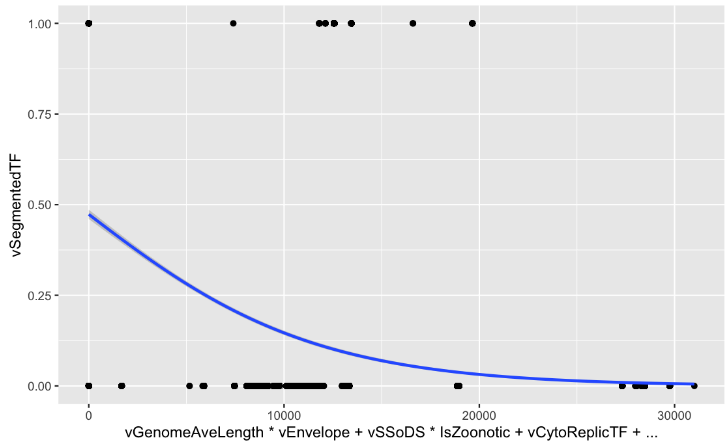
```
Call:
glm(formula = vSegmentedTF ~ vGenomeAveLength * vEnvelope + vSSoDS *
    IsZoonotic + vCytoReplicTF + vEnvelope, family = "binomial",
    data = df_hostremoved)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.287e+01  1.991e+02  -0.115    0.909
vGenomeAveLength           5.069e-04  5.337e-04   0.950    0.342
vEnvelope                  2.220e+01  3.511e+02   0.063    0.950
vSSoDS                    -2.170e+01  3.511e+02  -0.062    0.951
IsZoonotic                 1.976e+01  3.152e+02   0.063    0.950
vCytoReplicTF              1.819e+01  1.988e+02   0.092    0.927
vGenomeAveLength:vEnvelope -4.702e-04  5.338e-04  -0.881    0.378
vSSoDS:IsZoonotic         -1.734e+01  3.152e+02  -0.055    0.956

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21234.0  on 19600  degrees of freedom
Residual deviance:  7027.1  on 19593  degrees of freedom
AIC: 7043.1

Number of Fisher Scoring iterations: 19
```

Model excluding host order.



```
Call:
glm(formula = vSegmentedTF ~ HostOrder + vGenomeAveLength * vEnvelope +
    vSSoDS * IsZoonotic + vCytoReplicTF + vEnvelope, family = "binomial",
    data = df_hostremoved)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.390e+01  3.147e+02  -0.076   0.9395
HostOrdercarnivora         1.789e+00  1.194e-01  14.977  < 2e-16 ***
HostOrderchiroptera       -2.242e+00  2.211e-01 -10.139  < 2e-16 ***
HostOrderdidelphimorphia  -2.064e+01  4.575e+03  -0.005   0.9964
HostOrderdiprotodontia    -1.123e+00  4.766e-01  -2.356   0.0184 *
HostOrderlagomorpha       -1.166e+00  7.568e-01  -1.541   0.1233
HostOrderperissodactyla   -8.081e-01  1.207e-01  -6.697 2.13e-11 ***
HostOrderprimates         -8.287e-01  8.274e-02 -10.016  < 2e-16 ***
HostOrderrodentia         -3.059e+00  2.746e-01 -11.137  < 2e-16 ***
vGenomeAveLength           5.211e-04  3.241e-04   1.608   0.1079
vEnvelope                  2.373e+01  5.625e+02   0.042   0.9664
vSSoDS                    -2.301e+01  5.626e+02  -0.041   0.9674
IsZoonotic                 2.154e+01  4.885e+02   0.044   0.9648
vCytoReplicTF              1.933e+01  3.146e+02   0.061   0.9510
vGenomeAveLength:vEnvelope -4.795e-04  3.242e-04  -1.479   0.1391
vSSoDS:IsZoonotic         -1.897e+01  4.885e+02  -0.039   0.9690
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21234  on 19600  degrees of freedom
Residual deviance:  5915  on 19585  degrees of freedom
AIC: 5947

Number of Fisher Scoring iterations: 20
```

Model including host order.



Visualization of fitted model, without Host Order.

References

den Boon, J. A., Diaz, A., & Ahlquist, P. (2010). Cytoplasmic viral replication complexes. Cell host & microbe, 8(1), 77–85. https://doi.org/10.1016/j.chom.2010.06.010

McDonald, S. M., Nelson, M. I., Turner, P. E., & Patton, J. T. (2016). Reassortment in segmented RNA viruses: mechanisms and outcomes. Nature Reviews Microbiology, 14(7), 448–460. https://doi.org/10.1038/nrmicro.2016.46

Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. Cellular and molecular life sciences : CMLS, 73(23), 4433–4448. https://doi.org/10.1007/s00018-016-2299-6

Valero-Rello, A., & Sanjuán, R. (2022). Enveloped viruses show increased propensity to cross-species transmission and zoonosis. Proceedings of the National Academy of Sciences, 119(50). https://doi.org/10.1073/pnas.2215600119